Detecção de Conglomerados Espaciais com Geometria Arbitrária

MARCELO A. COSTA[1]
LUCIANO R. SCHRERRER[2]
RENATO M. ASSUNÇÃO[3]

Palavras-chave
estatística espacial - conglomerado - verossimilhança

Resumo

A detecção de conglomerados espaciais tem como objetivo a delimitação de uma região geográfica na qual a hipótese de ocorrência aleatória de um determinado evento pontual é rejeitada. Tal informação é de extrema relevância em estudos epidemiológicos. Este artigo apresenta um método de detecção de conglomerados espaciais no qual a estrutura de vizinhança espacial é agregada ao processo de crescimento e busca de conglomerados, possibilitando a detecção de conglomerados de geometria arbitrária. Os métodos tradicionais de varredura espacial restringem à geometria de busca a conglomerados de geometria circular, resultando em uma detecção parcial ou superestimação do conglomerado. Restrições durante o processo de crescimento são sugeridas para evitar conglomerados de tamanho excessivo e geometria muito irregular. Uma avaliação do poder de detecção do método para conglomerados com geometria arbitrária é realizada utilizando dados simulados. Resultados de detecção de conglomerados espaciais em dados de crimes são apresentados para a região de Belo Horizonte.

1. Introdução

Estudos de detecção de conglomerados espaciais são procedimentos importantes na área de vigilância em saúde pública. O diagnóstico preciso sobre a característica aleatória ou não de um determinado evento espacial como, por exemplo, uma doença contagiosa, e a delimitação da região geográfica de ocorrência possibilitam aos órgãos competentes a elaboração de políticas eficientes de controle e combate. Como resultado, procura-se identificar áreas geográficas com um risco significativamente elevado sem o conhecimento, em princípio, de quais e quantas áreas são, caracterizando um teste genérico de conglomerado.

Um conglomerado pode ser definido como um conjunto de áreas que apresentam um risco significativamente elevado quando considerada a hipótese nula (H0) de que os eventos são gerados aleatoriamente sobre a região de estudo. Um conglomerado pode ser caracterizado como temporal, espacial ou espaço-temporal, dependendo da variável (espaço e/ou tempo) sobre a qual é realizada a análise de contagem dos eventos. Especificamente, o presente trabalho abrange a detecção de conglomerados espaciais.

Diversas abordagens são apresentadas para a delimitação de conglomerados. Métodos computacionais puramente gráficos identificam regiões críticas a partir de sobreposição de círculos,

[1] azevedo@est.ufmg.br
[2] luscherrer@hotmail.com
[3] assuncao@est.ufmg.br

mas não fornecem uma medida de significância da região delineada [Bes91, OCC+88]. Neste contexto, o Método de varredura espacial scan proposto por Kulldorff [Kull97] tem sido amplamente utilizado em virtude do poder de detecção [KTP03, CA05] e da capacidade de atribuir um nível de significância à estatística de teste via simulação Monte Carlo, reduzindo o erro do tipo I. Entretanto, em sua formulação original, o método é condicionado à busca de conglomerados que apresentam geometria circular. Tal característica reduz substancialmente o custo computacional do método uma vez que uma busca exaustiva sobre todos os possíveis candidatos a cluster em uma área subdividida em n subáreas representa uma varredura sobre 2^{n-1} candidatos. Apesar da vantagem da eficiência computacional, o método apresenta limitações quando o conglomerado real passa a apresentar uma geometria irregular, detectando nenhuma ou pequenas áreas do mesmo. A principal vantagem da detecção de conglomerados de geometria arbitrária consiste na informação de dispersão e delimitação aproximada da região crítica que é incorporada ao resultado final. Como exemplo, suponha a existência de uma determinada formação geográfica como um rio, lago, vale, rodovia, córrego, a geometria encontrada pode ser comparada com tais formações e caso haja evidência de similaridade de geometria a formação geográfica em questão pode estar fortemente associada à formação ou disseminação do evento em estudo: doença, crime, etc. O tratamento da irregularidade do conglomerado tem sido abordado a partir de heurísticas computacionais, como o método de Simulated Annealing [DA04] ou delimitando uma região circular de tamanho fixo, menor que a região de estudo, e realizando uma busca exaustiva nas áreas contidas em seu interior [TT05]. Sob a suposição de que as regiões que definem o conglomerado compartilham fronteira geográfica, foi proposto o método de árvore dinâmica [ACT+06] que promove o crescimento de conglomerados agregando as áreas vizinhas que favorecem a maximização da verossimilhança do conglomerado.

Neste trabalho, o método originalmente proposto de crescimento dinâmico de conglomerados (dMST - dynamic Minimum Spanning Tree) é avaliado. A generalização da geometria resulta na identificação de conglomerados de tamanho elevado e muito irregular quando comparado com dados simulados. Com a finalidade de minimizar esses efeitos, são propostas restrições sob a verossimilhança das vizinhanças bem como sobre o tamanho mínimo do conglomerado a ser detectado. Os procedimentos propostos possuem baixa complexidade computacional e são capazes de promover a detecção de conglomerados mais compactos.

2. O Método de Varredura Scan

Veja uma região geográfica delimitada, subdividida em n subáreas, sendo associada a cada sub-área o número observado de casos yi e o número total de pessoas em risco na área, Ni. Como exemplo, pode-se definir uma região geográfica como um município subdivido em bairros (subáreas).

Cada bairro possui uma respectiva população (Ni) e uma determinada contagem relativa a um evento específico (yi) como número de indivíduos infectados por determinada epidemia ou número de indivíduos assaltados em um determinado período de tempo ou número de homicídios ocorridos no bairro em determinado ano, etc. Sob a hipótese nula de aleatoriedade ou ausência de conglomerados, o número esperado de casos na i-ésima área pode ser modelado por uma variável aleatória de Poisson e é independente das demais áreas sendo o número esperado de casos na área proporcional à população residente na mesma: $H_0 : y_i \sim Poisson(E_i = \lambda N_i),\ \mathrm{E}[y_i] = \lambda \mathrm{N_i}$ ***, onde a taxa estimada de ocorrência de casos é calculada como:*** $\hat{\lambda} = C/M$ ***, onde*** $C = \sum_i y_i$ ***e*** $M = \sum_i N_i$ ***.***

Seja Z o conjunto das áreas z candidatas a formarem um conglomerado. Se não for imposta nenhuma restrição espacial, o conjunto Z possui 2^{n-1} elementos. O vetor de parâmetros do método de máxima verossimilhança para o método scan é definido pela área candidata z, a probabilidade de que um indivíduo

em z seja um caso (p), e a probabilidade de um indivíduo fora de z seja um caso (r). Sob a hipótese nula: p=r e sob a hipótese alternativa: p>r. Definindo n_z como a população em z, c_z o número de casos em z, $\hat{p} = c_z / n_z$ e $\hat{r} = (C - c_z)/(M - n_z)$, a função de verossimilhança do candidato a conglomerado é definida por:

$$L(z, p_z, r_z) = \sup_{z \in Z, p>r} p^{c_z}(1-p)^{(n_z - c_z)} r^{(C-c_z)}(1-r)^{(M-n_z-C+c_z)} \quad (1)$$

referente ao modelo de Bernoulli, para todo $z \in Z$ e que representa o produto da probabilidade da ocorrência de c_z casos dentro do cluster z versus a probabilidade de ocorrência de $C - c_z$ casos fora sob a restrição de que $p_{(z)} > r_{(z)}$. Ao conglomerado verossímil é atribuída uma estatística baseada na razão de verossimilhança: $\kappa = L(\hat{z}, \hat{p}_z, \hat{r}_z)/L_0$, onde $L_0 = C^C (M-C)^{M-C} / M^M$.

Em sua proposta inicial [Kull97], o conjunto Z representa círculos de raio r arbitrário centrados em cada um dos n centróides das subáreas. A varredura sobre o conjunto circular de áreas é realizada a partir de um processo iterativo no qual um círculo é posicionado no centróide da primeira subárea. O raio é inicialmente dimensionado de forma que apenas a respectiva área seja contemplada pelo conglomerado. Uma vez que a estatística κ seja calculada, o raio é incrementado de modo a abranger a subárea mais próxima. O procedimento de aumento sucessivo do raio e cálculo da estatística κ é repetido até que o conglomerado alcance um tamanho máximo de população ou de subáreas. O processo é então interrompido e repetido a partir de uma nova subárea até que sejam formados círculos a partir de todos os centróides. O círculo que obteve o maior valor para a estatística κ é armazenado e prossegue-se no cálculo do nível descritivo da estatística. A restrição circular para a geometria de busca reduz significativamente o número de candidatos a conglomerados e, conseqüentemente, o custo computacional.

A Figura 1 apresenta trecho do código C do loop de varredura dos conglomerados circulares utilizando o modelo de verossimilhança de Bernoulli. Para simplificação do processo de incremento do raio, os índices dos centróides (1...n) são representados em uma forma matricial (m_index.data[k][i], dimensão: n x n) onde cada k-ésima linha da matriz contém os índices de todos os centróides em ordem crescente de distância euclidiana em relação ao centróide da sub-área k.

```
lambda = -1;                    /* Valor inicial da estatística de
teste */

for(k=0;k<centroides;k++)  /* Varredura sobre todos os centróides */
{
   cz = 0;
   nz = 0;

   for(i=0;i<(centroides-1);i++){   /* aumento progressive do raio */
      cont = (int) m_index.data[k][i];
      cz = cz + casos.data[cont][0];         /* contagem dos casos */
      nz = nz + pop.data[cont][0];        /* contagem da população */
      p = cz/nz;           /* probabilidade dentro do conglomerado */
      r = (C-cz)/(N-nz);   /* probabilidade fora do conglomerado */

      if(p<=r)             /* Verifica a restrição de prob. */
         L = Lo;

      else                 /* Calcula a verossimilhança (Bernoulli)*/
         L = (cz*log(p)) + ((nz-cz)*log(1-p)) + ((C-cz)*log(r)) +
         ((N-nz-C+cz)*log(1-r));
         aux = L-Lo;

       if(aux > lambda){/* Armazena o conglomerado mais verossímil */
         lambda = aux;
         lambda_x = k;
         lambda_y = i;
         lambda_p = p;
         lambda_r = r;
       }

       if(nz>Max_pop)   /*Quebra o Loop se a população interna é
                          maior que 20% da total*/
       break;
     }
}
```

Figura 1. Trecho do código C responsável pela varredura dos conglomerados circulares.

O algoritmo de varredura sempre retorna o conglomerado de máxima verossimilhança. Para se testar a hipótese de que o conglomerado detectado é realmente significativo é necessária a distribuição da estatística κ, cuja solução analítica é de difícil obtenção. Contudo, a sua distribuição empírica condicionada ao número total de casos, sob a hipótese nula (H_0) é obtida via simulação Monte Carlo a partir dos seguintes passos:

1. *Gera-se S conjuntos independentes de vetores de casos, cuja soma dos elementos de cada vetor seja C, a partir de realizações de uma distribuição multinomial proporcional à população de cada área. Calcula-se a estatística κ para cada conjunto: ($\kappa_1, \ldots, \kappa_S$);*
2. *Ordena-se os valores de κ. Se o valor obtido com o conjunto de dados original estiver entre os maiores 100(1-α)%, rejeita-se H_0 ao nível de significância α.*
3. *Caso H_0 tenha sido rejeitada, a área $\hat{z}$ associada é o conglomerado mais verossímil.*

Neste contexto, a rejeição da hipótese nula de aleatoriedade indica que existe evidência de que a ocorrência do conglomerado detectado não é meramente aleatória.

3. O Algoritmo de Construção de Conglomerados com Geometria Irregular

Seja a região de interesse definida por subáreas que compartilham fronteira geográfica de forma que, para uma particular subárea i, exista pelo menos uma outra subárea j que possui fronteira comum. Pode-se expressar essa informação sob a forma de um grafo interconectando os centróides das subáreas aos seus vizinhos, conforme ilustra a Figura 2-A.

Uma árvore geradora mínima de um grafo representa um subgrafo interconectando todas as arestas, mas cujo caminho entre dois nodos i e j seja único, de tal forma que se uma aresta da árvore geradora mínima for removida obtêm-se dois subgrafos não conectados. A Figura 2-B ilustra uma árvore geradora mínima.

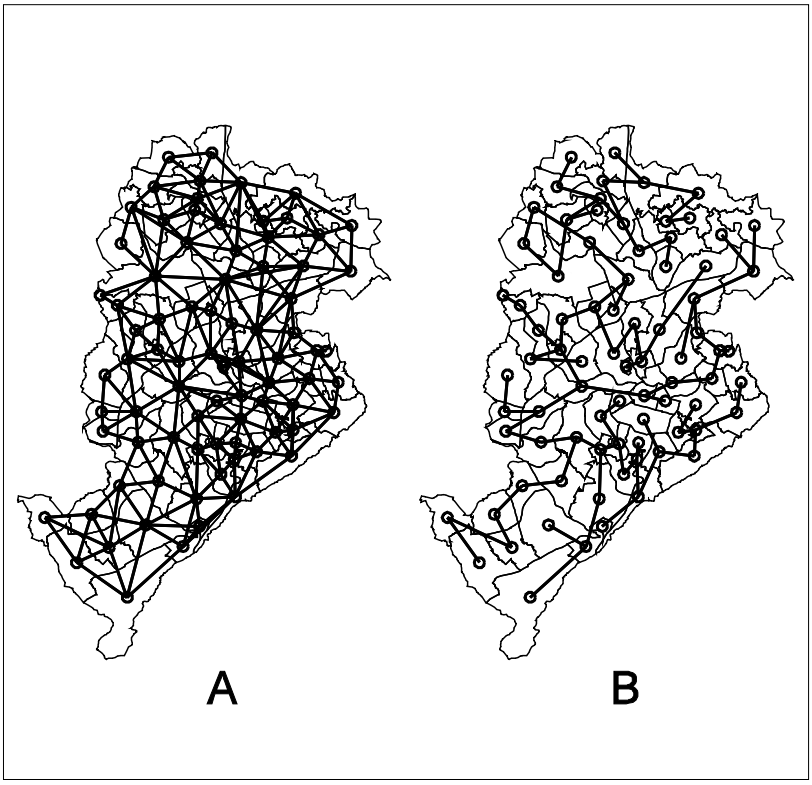


Figura 2. Mapa de Belo Horizonte subdividido em unidades administrativas interconectadas por um grafo de arestas (A) e o grafo da árvore geradora mínima (B)

O algoritmo para a construção de geometrias arbitrárias, denominado dMST (dynamic Minimum Spanning tree), tem como objetivo a construção de árvores geradoras mínimas na qual o custo de agregação de uma área à árvore está associada à verossimilhança da árvore resultante. O algoritmo de crescimento da árvore geradora mínima utilizando a equação de verossimilhança é descrito a seguir:

1. *Partindo de cada subárea, calcule a verossimilhança de acordo com a Equação 1 considerando cada um de seus vizinhos como possíveis candidatos a incorporar o conglomerado;*
2. *Inclua na árvore o vizinho que resulta na maior verossimilhança;*
3. *Defina os vizinhos da nova árvore;*

Retorne à etapa 2 e repita o procedimento até que todas as subáreas estejam incluídas na árvore geradora mínima ou até que a árvore alcance um tamanho máximo predefinido.

```
for(cntarea=0;cntarea<NumAreas;cntarea++){/* Loop de crescimento da */
                                          /* arvore                 */
   clustScan.tamanho = 1;                              /*  Inicializaçao
da estrutura de varredura */
   clustScan.indices[0] = cntarea;

   /* Calcula a verossimilhanca da arvore de varredura */
   clustScan.veross = VerossimilhancaCluster(&clustScan,VetAreas);

   /* Verifica limite de areas na arvore */
   while(clustScan.tamanho<LimiteAreas){
       /* Calcula os vizinhos da arvore de varredura */
      VizinhosCluster(&vizinClust,&clustScan,VetArestas);

      /* Promove o crescimento da arvore e verifica se */
       /* a arvore resultante é a mais verossimil       */
       if(AgregaAreaCluster(&clustScan,&vizinClust)>bestClust.veross){
         CopiaCluster(&clustScan,&bestClust);   /* Armazena a arvore */
         }
     }
}
```

Figura 3. Trecho do código C++ responsável pelo crescimento da árvore de varredura.

Em sua proposta original, demonstrada na Figura 3, o critério de parada do algoritmo de construção de árvores é o tamanho máximo especificado pelo usuário (LimiteAreas) e, para uma região com n subáreas, são geradas n árvores. Durante o crescimento das árvores, o método armazena a estrutura (subárvore) de máxima verossimilhança. Em seqüência, o método de simulação de Monte Carlo é utilizado para o cálculo do nível descritivo associado à estatística da razão de verossimilhança κ*, sob* H_0*. A análise final é semelhante à descrita para o conglomerado circular.*

4. Metodologia

Uma análise de desempenho do método dMST é proposta a partir de dados simulados e dados reais. A região de interesse é representada pela região metropolitana de Belo Horizonte subdividida em bairros.

Os dados simulados foram gerados a partir de três cenários distintos para os conglomerados, apresentados na Figura 4. A população de interesse em cada bairro foi obtida a partir do censo do ano de 2000. No primeiro cenário, especificou-se um conglomerado com geometria circular constituído

por 13 bairros. No segundo cenário, especificou-se um conglomerado com geometria estrela constituída por 12 bairros e no terceiro cenário, utilizou-se uma geometria retilínea definida por 6 bairros. O número total de casos na região de estudo é de 420, sendo distribuídos de acordo com uma distribuição multinomial na qual as probabilidades referentes aos bairros do conglomerado foram ajustadas a partir da especificação de um risco relativo, favorecendo a rejeição da hipótese nula com probabilidade 0.999 [KTP03]. Uma vez definidos os parâmetros de simulação, foram geradas 10.000 simulações para cada cenário onde, para cada simulação, foram distribuídos 420 casos entre os bairros. Em seguida avaliou-se o poder de detecção do método scan circular, dMST e mais duas variações propostas:

1. *$dMST_2$: Método dMST com parada prematura. Nesta abordagem a árvore irá crescer enquanto existir algum vizinho que, ao ser acrescentado, resulta em uma árvore com verossimilhança maior que a árvore anterior, caso contrário o crescimento é interrompido e uma nova árvore é gerada a partir das demais áreas;*
2. *$dMST_3$: Método dMST com parada prematura, busca suavizada e tamanho mínimo. Semelhante à abordagem anterior, inicialmente a árvore cresce até atingir um tamanho mínimo. Em seqüência, é agregado à árvore o vizinho que proporciona o menor crescimento da verossimilhança dentre todos os vizinhos capazes de maximizar a verossimilhança em relação à árvore anterior. Caso nenhum vizinho proporcione o aumento da verossimilhança, o método é interrompido e novas árvores são geradas a partir das demais áreas.*

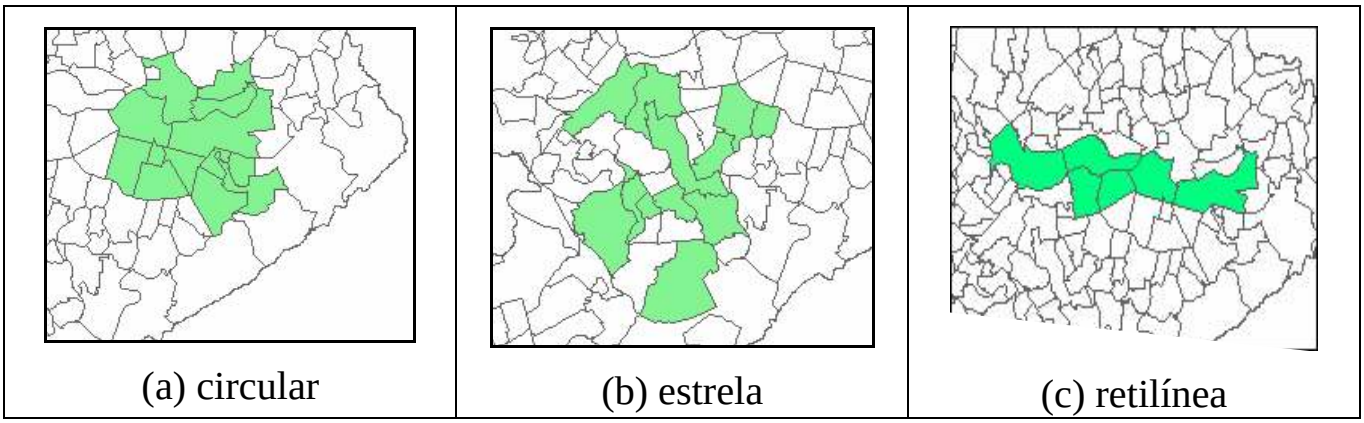


Figura 4. Cenários de conglomerados para dados simulados

Para avaliar o desempenho das metodologias em situações reais, aplicou-se a análise de conglomerados para os dados de homicídios relatados nos anos 2000 e 2001.

5. Resultados

A partir das 10.000 simulações avaliou-se o poder de detecção dos métodos scan, dMST, $dMST_2$ e $dMST_3$. Os resultados obtidos para o método scan são mostrados na Figura 5, na qual a escala dos eixos foram previamente padronizada: 0-60 (eixo x) e 0-5.000 (eixo y) para os gráficos de contagem de tamanho detectado e 0-3.000 (eixo y) para os gráficos de interseção sendo a escala do eixo x definido em função do cenário: 0-13 (circular), 0-12 (estrela) e 0-6 (retilíneo). Para o cenário circular, o método scan apresentou um bom desempenho, mas com um número elevado de conglomerados detectados sem interseção com o conglomerado real (2.225). Para os cenários estrela e retilíneo, ocorreu uma queda de desempenho e uma maior irregularidade na distribuição das interseções. O cenário retilíneo apresentou o maior número de conglomerados detectados sem interseção (2.871). Em todos os cenários, o método detectou com maior freqüência, conglomerados de tamanho unitário.

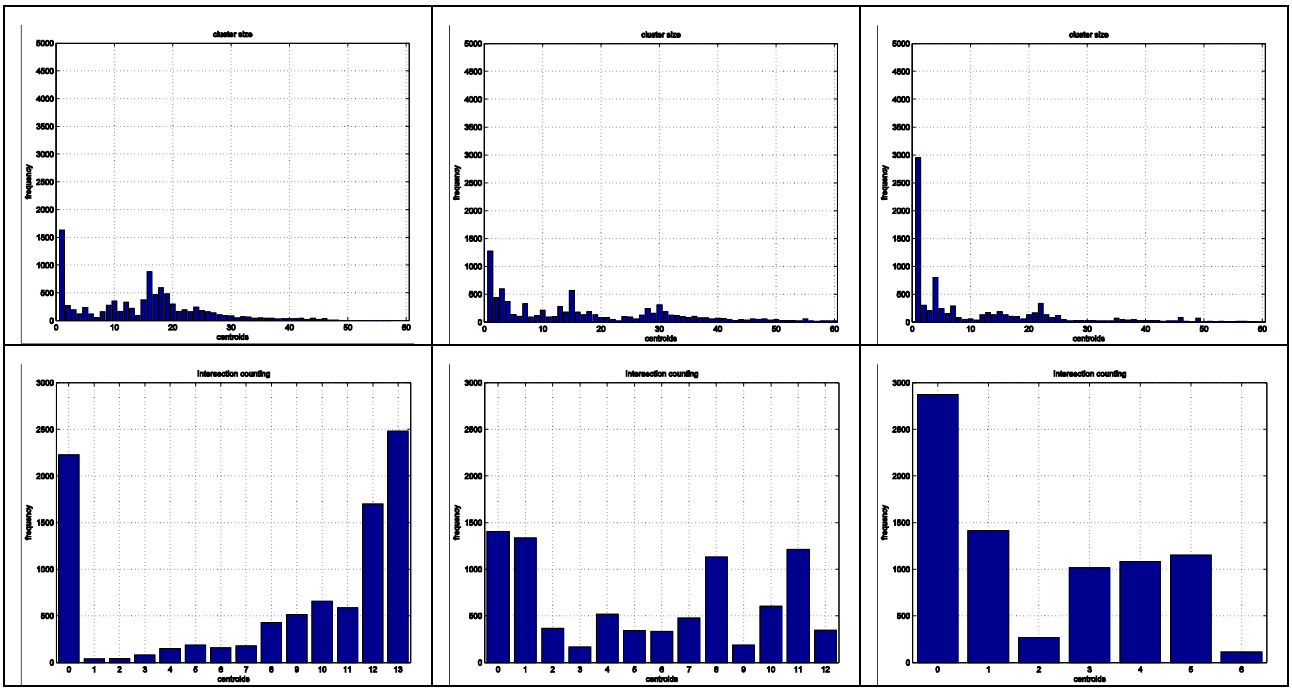


Figura 5. Distribuição do tamanho do conglomerado encontrado (linha 1) e distribuição da contagem da interseção entre o conglomerado encontrado e o conglomerados real (linha 2), pelo método scan para os cenários circular (coluna 1), estrela (coluna 2) e retilíneo (coluna 3).

As Figuras 6 e 7 apresentam os resultados para os métodos dMST, $dMST_2$ e $dMST_3$. O método dMST detectou conglomerados com dimensões elevadas, concentrados próximas do limite de tamanho máximo (60 bairros) em todos os cenários. As distribuições das interseções são bem regulares com médias inferiores ao tamanho real do conglomerado mais, próximas do mesmo (ver Figura 7). Os métodos $dMST_2$ e $dMST_3$ possibilitaram uma redução da média da distribuição do tamanho detectado. O método $dMST_2$ gerou uma contagem maior de elementos sem interseção com o conglomerado real, característica que foi minimizada pelo método $dMST_3$. Por outro lado, o método $dMST_3$ detectou uma quantidade maior de conglomerados com interseções de 1 e 2 bairros em relação ao conglomerado real para os cenários circular e estrela. A Tabela 1 apresenta a contagem entre as 10.000 simulações realizadas, que foram efetivamente consideradas, uma vez que foram consideradas apenas as simulações nas quais a hipótese nula era rejeitada. Destas, também são apresentadas as contagens de interseção nula, ou seja, os casos nos quais o conglomerado detectado não apresenta nenhuma interseção com o conglomerado real. A partir desta Tabela pode-se comparar os métodos em relação à detecção efetiva, por exemplo, apesar do método scan não rejeitar 9.423 simulações (circular) ocorreram 2.225 casos sem interseção, por outro lado o método $dMST_3$ não rejeitou 8.002 simulações, mas ocorreu um número menor de simulações sem interseção, 517. No geral, avaliando-se a capacidade de detecção, os métodos de detecção de geometria arbitrária apresentaram resultados superiores ao método scan para os cenários circular e retilíneo e inferiores para o cenário estrela. A avaliação de desempenho das metodologias propostas da detecção em cenários simulados fornece medidas de sensibilidade quando os mesmos são aplicados a cenários reais. Neste contexto, o número de casos na região de estudo foi fixado em 420 para facilitar a comparação dos resultados de simulação com os resultados obtidos a partir dos dados de Homicídios em Belo Horizonte durante o ano de 2000.

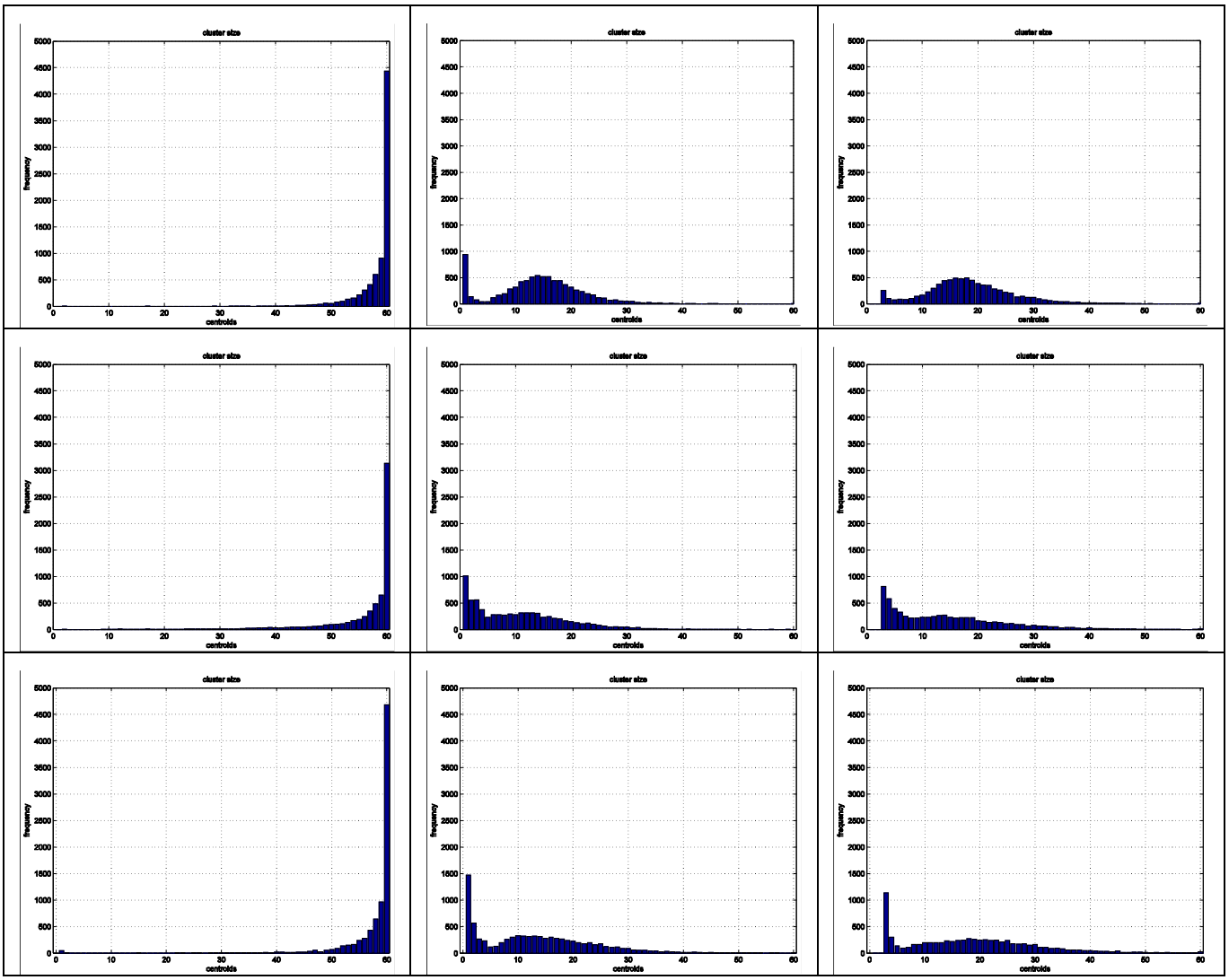


Figura 6. Distribuição do tamanho do conglomerado encontrado pelo método dMST (coluna 1), $dMST_2$ (coluna 2) e $dMST_3$ (coluna 3) para os cenários circular (linha 1), estrela (linha 2) e retilíneo (linha 3).

A Figura 8 apresenta os resultados para os dados de Homicídio em Belo Horizonte durante o ano de 2000. O conglomerado detectado pelo método scan abrange 16 bairros, sendo que em alguns bairros não são observados casos. O método dMST obteve o maior conglomerado (48 bairros) e, visualmente, o conglomerado resultante é uma interligação de vários subconglomerados. O método $dMST_2$ identificou um conglomerado de tamanho 1 e o método $dMST_3$ obteve um conglomerado de tamanho 8 constituído pelo conglomerado detectado pelo método $dMST_2$, parte do conglomerado detectado pelo método dMST e um bairro sem contagem interligando os mesmos.

Apesar da discrepância em relação à irregularidade da geometria do conglomerado detectado pelo método dMST e do número de áreas sem casos detectadas pelo método scan, é evidente a formação de um cluster a partir da interseção das áreas encontradas por cada método, à exceção do método $dMST_2$. Com base nos resultados de simulação e na característica de interseção dos resultados, existe evidência de que o conglomerado apresentado pelo método $dMST_3$ pode ser considerado como o resultado final do processo de busca.

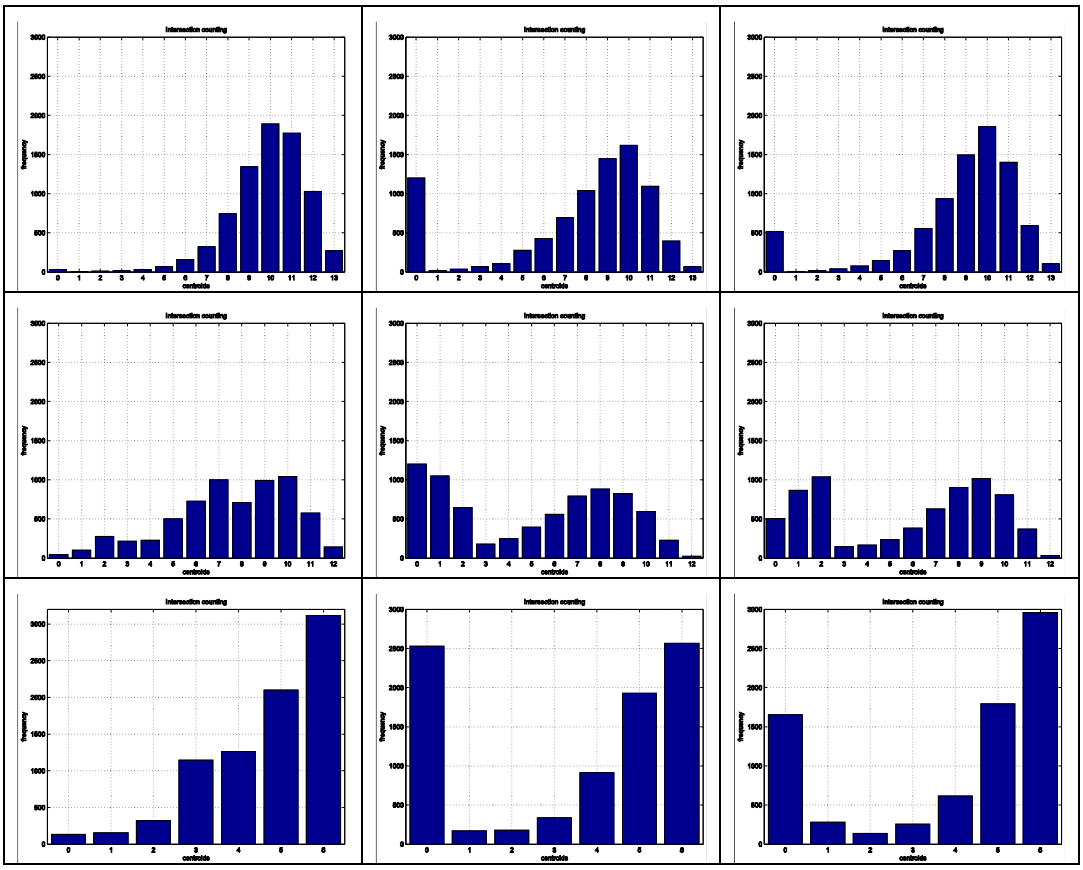
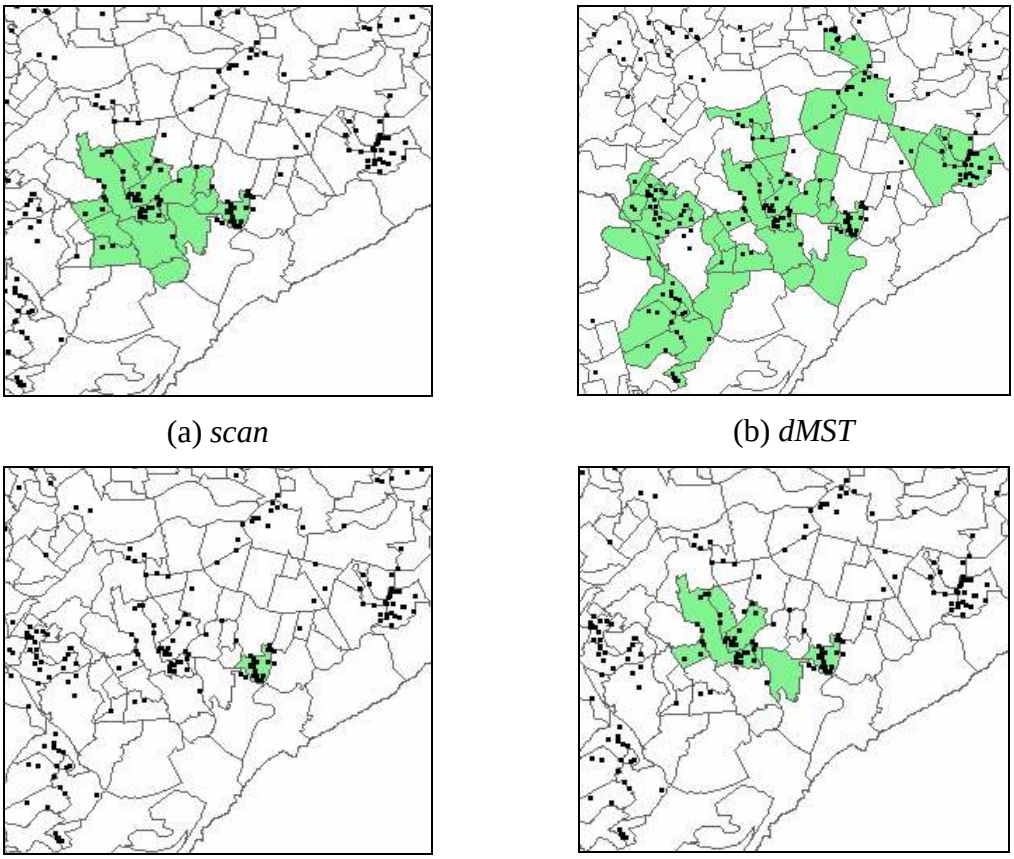


Figura 7. Distribuição da contagem da interseção entre o conglomerado encontrado e o conglomerado real obtido pelo método dMST (coluna 1), $dMST_2$ (coluna 2) e $dMST_3$ (coluna 3) para os cenários circular (linha 1), estrela (linha 2) e retilíneo.

A Figura 9 apresenta os resultados de identificação de conglomerado utilizando os dados de homicídio do ano de 2001. Visualmente os métodos $dMST_2$ e $dMST_3$ detectaram o mesmo conglomerado, sendo que o mesmo está contido no conglomerado detectado pelo método scan. Uma vez que não existe nenhuma subárea com contagem nula de casos, o processo de busca de detecção é finalizado. Para os conglomerados detectados pelos métodos mencionados, foram obtidos p-valores, via simulação de Monte Carlo, inferiores a 0,05 (5%)

Tabela 1. Contagem do número de simulações, em 10.000, nas quais a hipótese nula foi rejeitada e, dentre essas, comparação com a contagem de simulações sem interseção entre o conglomerado encontrado e o conglomerado real para os cenários: circular, estrela e retilíneo.

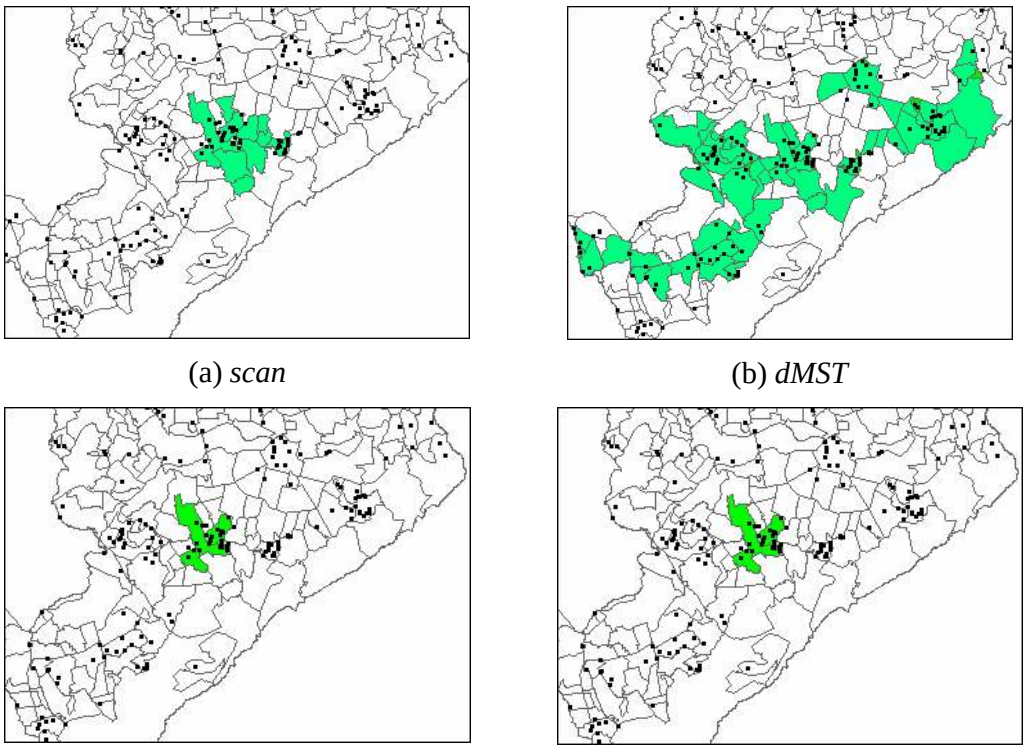
Método	Característica	Circular	Estrela	Retilínio
scan	*p*-valor < 0.5	9423	8405	7907
	sem interseção	2225	1403	2871
dMST	*p*-valor < 0.5	7669	6549	8224
	sem interseção	29	45	133
$dMST_2$	*p*-valor < 0.5	8476	7605	8618
	sem interseção	1202	1199	2529
$dMST_3$	*p*-valor < 0.5	8002	7082	7695
	sem interseção	517	505	1655



(a) *scan* (b) *dMST*

(c) $dMST_2$ (d) $dMST_3$

Figura 8. Conglomerados identificados para os dados de Homicídios em Belo Horizonte durante o ano de 2000.



(a) *scan* (b) *dMST*

(c) $dMST_2$ (d) $dMST_3$

Figura 9. Conglomerados identificados para os dados de Homicídios em Belo Horizonte durante o ano de 2001.

8. Discussões e Conclusões

O principal objetivo do processo de vigilância espacial é a identificação e a classificação de um conglomerado. É necessário verificar se existe evidência da ocorrência do conglomerado ao acaso. Este fato é determinado pelo resultado do p-valor obtido via simulação Monte Carlo. Em resumo, procede-se com o método de varredura, identificando o conglomerado mais verossímil e, em seguida, avalia-se a ocorrência ao acaso do mesmo. Caso o p-valor seja menor que 5%, valor previamente definido pelo usuário, a hipótese nula de ocorrência ao acaso é rejeitada e o conglomerado passa a ser avaliado como significativo.

A partir dos dados simulados foram levantados alguns aspectos em relação às características dos conglomerados detectados por cada método. Para os cenários simulados, o método scan detecta conglomerados com elevada intensidade de bairros sem casos, provavelmente devido à própria restrição da geometria. Este fato também é observado no método dMST com geometrias arbitrárias. Neste caso, há evidência de que esta característica está relacionada ao crescimento exagerado da árvore uma vez que o método $dMST_3$ minimiza este efeito e o método $dMST_2$ praticamente o elimina.

Do ponto de vista de freqüência de detecção do conglomerado real nas simulações, seja na totalidade ou na parcialidade, ambos os métodos apresentam desempenhos similares. O método scan apresenta uma baixa taxa de rejeição da hipótese nula mas uma alta taxa de conglomerados sem interseção. Os métodos dMST, $dMST_2$ e $dMST_3$ apresentam alta taxa de rejeição da hipótese nula mas, baixa taxa de não-interseção.

Na aplicação aos dados reais, a análise da interseção de todos os métodos indica uma concentração anormal de casos em um único bairro. Por outro lado, um segundo conglomerado pode ser identificado a partir das interseções dos bairros detectados pelos métodos scan, dMST e $dMST_3$.

A partir dos resultados obtidos pode-se concluir que, para as bases de dados estudadas, incorporar a estrutura de vizinhança torna o método scan mais focado permitindo a redução do número de interseções nulas e do número de regiões sem casos no conglomerado final. A metodologia também permite obter, além de um valor para a estatística de teste e um p-valor associado, a geometria do conglomerado. Tal informação pode ser comparada com outras informações espaciais, como a estrutura do relevo da região, na busca de possíveis causas de ocorrência do fenômeno em estudo.

Spatial Cluster Detection with Arbitrary Shape

Keywords:
Spatial statistics cluster - likelihood

Abstract
Spatial cluster detection aims at detecting a particular region in which the hypothesis of random occurrence of an event is rejected. This information is of extreme relevance in epidemiology studies. This article presents a cluster detection method that aggregates the spatial neighbor structure into the cluster growing process. The procedure allows the detection of arbitrarily shaped clusters. Standard scan spatial statistics confine the cluster geometry shape to circular shaped clusters, resulting in partial or over sized clusters. Restrictions during the growth process are suggested in order to prevent over sized clusters with odd geometries. An evaluation of the method's performance is provided through simulation studies. Results of cluster detection in crime data are presented for Belo Horizonte city.

Referências Bibliográficas

[ACT+06] Assunção, R., Costa, M., Tavares, A., Ferreira, S. (2006), Fast detection of arbitrary shaped disease clusters. Statistics in Medicine. Forthcoming

[BN91] Besag, J. and Newell, J. (1991), The detection of clusters in rare diseases. Journal of the Royal Statistic Society A, vol. 154, pages 143-155.

[CA05] Costa, M. A., Assunção, R. M. (2005), A fair comparison between the spatial scan and the Besag-Newell disease clustering tests. Environmental and Ecological Statistics, vol. 12, pages 297-315.

[DA04] Duczmal, L. and Assunção, R. (2004), A simulated annealing strategy for the detection or arbitrarily shaped spatial clusters. Computational Statistics & Data Analysis, vol. 45, pages 269-286.

[KTP03] Kulldorff, M., Tango, T., Park, P. J., (2003), Power comparisons for disease clustering tests. Computational Statistics & Data Analysis, vol. 42, pages 665-684.

[Kull97] Kulldorff, M. (1997), A Spatial Scan Statistics. Commun. Statist. Theory and Methods., vol. 26, pages 1481-1496.

[OCC+88] Openshaw, S., Craft, A. W., Charlton, M. and Birch, J. M. (1988), Investigation of leukaemia clusters by use of a geographical analysis machine. Lancet i, pages 272-273

[TT05] Tango, T., Takahashi, K. (2005), A flexibly shaped spatial scan statistic for detecting clusters. International Journal of Health Geographics, 4:11.

Agradecimentos

Os autores agradecem à FAPEMIG e à Pró-Reitoria de Pesquisa da UFMG pelo apoio financeiro e à Polícia Militar de Belo Horizonte pelo auxílio e suporte à base de dados.

Sobre os Autores:

Marcelo Azevedo Costa

Professor Adjunto do Departamento de Estatística da Universidade Federal de Minas Gerais
Doutor em Engenharia Elétrica pela UFMG, 2002.
Áreas de interesse: estatística espacial, inteligência computacional, aprendizado de máquina.

Luciano Rios Schrerrer

Mestrando em Estatística Programa de Pós-Graduação em Estatística UFMG

Renato Martins Assunção

Professor Adjunto do Departamento de Estatística da Universidade Federal de Minas Gerais.
Ph.D. in Statistics - University of Washington, Seattle, 1994
Áreas de interesse: estatística espacial